Recitation 5: November 23

Lecturer: Regev Schweiger *Scribe: Yishay Mansour*

5.1 Learning a Threshold

We will show a slightly different proof than the one we have seen in class. Let X be the interval $[0, 1]$, and let $\mathcal{H}$ be a concept class over X:

$$\mathcal{H} = \{c_\theta \mid 0 \leq \theta \leq 1\}$$

where

$$c_\theta = \begin{cases} 1 & x > \theta \\ 0 & \text{otherwise} \end{cases}$$

Assume our data is generated by unknown distribution D, and the correct concept is $c_t \in \mathcal{H}$ (corresponds to a threshold t).

Let A be the algorithm the operates as follows: It examines m samples $\langle x_i, y_i \rangle$, where $x_i \sim D$ and $y_i = c_t(x_i)$. It then calculates a lower bound for the correct t:

$$b(\langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle) = \max\{x_i | y_i = 0\}$$

If there are no samples for which $y_i = 0$, set $b = 0$. Finally, the algorithm returns c_b as the hypothesis.

Denote $D[u, v] = \Pr_{x \in D}[x \in [u, v]]$, and similarly, $D(u, v]$. We would like to prove, that in probability at least $1 - \delta$, $\Pr_{x \sim D}[c_t(x) \neq c_b(x)] \leq \varepsilon$. Note that, by definition, $b(\langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle) \leq t$. We will therefore try to show:

$$\Pr_{x_1 \sim D, \ldots, x_m \sim D}[D\left(b(\langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle), t\right] > \varepsilon] \leq \delta$$

We shall define a parameter that does not depend on the sample. First, we note that, if $D[0, t] < \varepsilon$, then the accuracy bound holds with confidence 1. We therefore assume $D[0, t] \geq \varepsilon$. Let β be the point which is ε close to t from below. Namely:

$$D\left[\beta, t\right] = \varepsilon$$

(More precisely, we choose the β such that $D[\beta, t] \geq \varepsilon$ and $D(\beta, t] \leq \varepsilon$. This distinction may occur in some discrete distributions). It is sufficient to show that with high probability $(1 - \delta)$, $b \in [\beta, t]$.

We have that:

$$\begin{aligned} & \Pr_{x_1 \sim D, \ldots, x_m \sim D} \left[b(\langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle) \notin [\beta, t] \right] \\ = & \prod_{i=1}^{m} \Pr_{x_i \sim D} \left[x_i \in [0, \beta] \cup x_i \in [t, 1] \right] \\ = & \prod_{i=1}^{m} (1 - \varepsilon) = (1 - \varepsilon)^m \leq e^{-m\varepsilon} \end{aligned}$$

Finally we get:

$$\begin{aligned} & \Pr_{x_1 \sim D, \ldots, x_m \sim D} \left[D\left(b, t\right] > \varepsilon \right] \\ = & \Pr_{x_1 \sim D, \ldots, x_m \sim D} \left[b \notin [\beta, t] \right] \\ \leq & \; e^{-m\varepsilon} \end{aligned}$$

We want to see when $e^{-m\varepsilon} < \delta$. A simple calculation shows that this happens when the sample size m satisfies: $m > \frac{1}{\varepsilon} ln \frac{1}{\delta}$.

5.2 Learning Rectangles - Sample Size

We give a proof that the concept class of rectangles discussed in class is PAC-learnable, and analyse the sample size.

In this section, we try to find what is a "sufficiently large" number of examples that is needed to learn a good hypothesis. For that, we will fix our accuracy and confidence parameters (ε, δ), and the strategy A. We will show that for any distribution D, we can assert sample size m that with high confidence (that is, probability at least $1 - \delta$), the returned rectangle R' from strategy A (i.e. the tightest fit rectangle) has an error of at most ε.

We will construct strips $T_1, .., T_4$ such that $\forall i \; D(T_i) \leq \frac{1}{4}\varepsilon$ (Figure 5.1). From the construction we can see that T_i is independent of the sample and of R'. Note that we cannot certainly find the T_i but we can be sure that such T_i exists. Let the strips $T_i' \subseteq T_i$ be the strip from the respective edge of R to the first sample.

We would want to have $\forall i \; T_i' \subseteq T_i$. If that is the case, we obtained our requirement since

$$\Pr[error] = \mathcal{D}(R \Delta R') = \mathcal{D}(\cup T_i') \leq \mathcal{D}(\cup T_i) \leq \varepsilon$$

From the construction, if there is at least one sampled point that resides in T_i it implies that $T_i' \subseteq T_i$. This is true, since the rectangle from strategy A must include all sampled positive

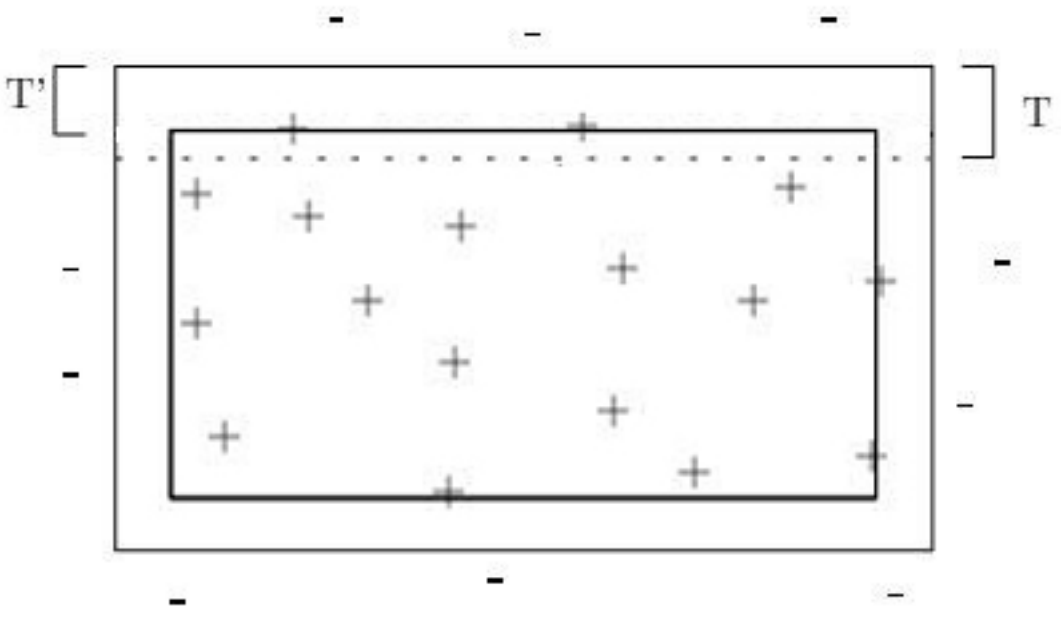

Figure 5.1: Adjusting strip size to have a weight of at most ε according to the real target function R.

points in R. To achieve that we can ask: what is the probability of sampling a bad event, that is, what is the probability that we didn't receive points from the sample data that are located on the constructive strips, i.e., T_i. Formally,

$$\Pr[\text{error} > \varepsilon] \leq \Pr[\exists i = 1..4 \; \forall \mathbf{x} \in S, \mathbf{x} \notin T_i]$$

By definition of T_i,

$$\Pr[x \notin T_i] = 1 - \frac{\varepsilon}{4}$$

Since our sample data is i.i.d from distribution D:

$$\Pr[\forall \mathbf{x} \in S, \mathbf{x} \notin T_1] = \left(1 - \frac{\varepsilon}{4}\right)^m$$

The same analysis holds on each T_i strip. Hence, we get that on the entire region the error would not exceed the sum of probabilities for each of the strips. That is,

$$\Pr[\text{error} > \varepsilon] \leq 4\left(1 - \frac{\varepsilon}{4}\right)^m$$

From the inequality $(1 - x) \leq e^{-x}$, we obtain:

$$\Pr[\text{error} > \varepsilon] \leq 4\left(1 - \frac{\varepsilon}{4}\right)^m \leq 4e^{-\frac{\varepsilon}{4}m} < \delta$$

That is, if we want to have accuracy ε and confidence of at least $1-\delta$, we have to choose the sample size m to satisfy:

$$4e^{-\frac{\varepsilon}{4}m} < \delta \Leftrightarrow m > \frac{4}{\varepsilon} ln \frac{4}{\delta}$$

For this strategy A, and for every small (ε, δ) we like, we got the sample size that is needed for having a good learner.

5.2.1 Remarks

1. The analysis holds for any fixed probability distribution $\mathcal{D}$, we only required that the sample points are i.i.d from distribution $\mathcal{D}$ to obtain our bound.

2. The minimal sample size $m(\varepsilon, \delta)$ behaves as we might expect. One might want to have better accuracy by decreasing ε or greater confidence by decreasing δ — our algorithm requires more examples to meet those requirements. There is a stronger dependence in ε.

3. The parameter ε gives the degree of accuracy that we want to achieve. It determines what is a good hypothesis for achieving a good approximation in respect to the target function. In our example, the accuracy determines which of our hypothesis rectangles are good enough in respect to the real rectangle target. We pay attention that the accuracy does not depend on the data distribution.

4. The parameter δ gives the degree of confidence on having a good learner. Meaning, how sure are we that we've reached that level of accuracy. This can be related on, how typical the given sample data reflects the true distribution. Again, it does not depend on the data distribution.

5. We might have cases as shown in Figure 5.2, where the distribution $\mathcal{D}$ gives large weights to particular regions of the plain, creating a distorted image of the rectangle. In any case, under those conditions, since the learner is tested on the same distribution $\mathcal{D}$, and this distribution has small error between R and R', the rectangle $\mathcal{R}'$ will be a good hypothesis (in respect to ε, δ).

6. The strategy A that we defined is efficient: In computational view, the only need is to search for the max and min points that defines our tightest-fit rectangle. In sample data size view, the number of examples that is required for achieving accuracy ε with confidence $1-\delta$ is polynomial in $\frac{1}{\varepsilon}$ and $ln\frac{1}{\delta}$.

7. In this example, as opposed to the Bayesian approach, we haven't been trying to model $\mathcal{D}$ or to guess which rectangle is more likely (prior). We have separated the distribution

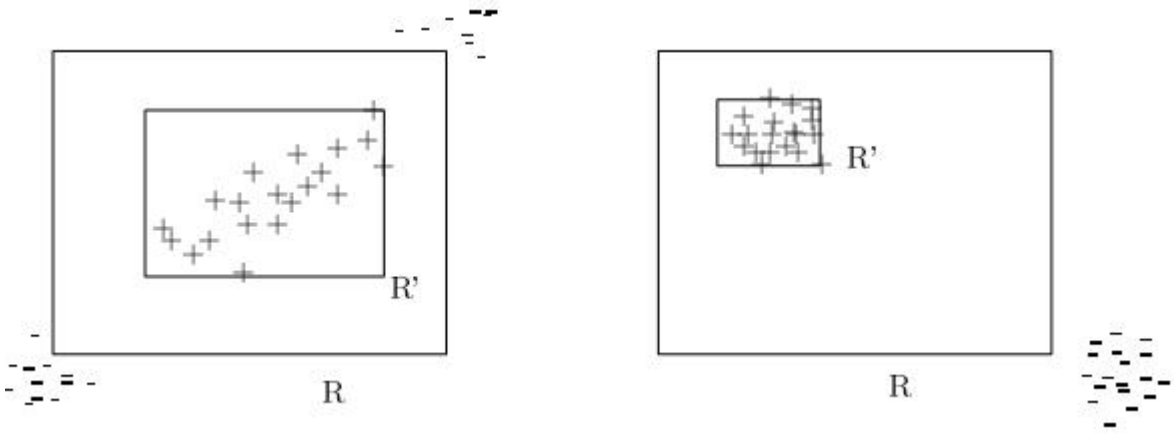

Figure 5.2: Two cases depending on the sample size

$\mathcal{D}$ from the target function (rectangle R), and directly try to predict hypothesis for this function.